SYSTEM AND METHOD FOR EFFICIENT REPRESENTATION OF DATA SET ADDRESSES IN A WEB CRAWLER

ABSTRACT OF THE DISCLOSURE

A web crawler stores fixed length representations of document addresses in a buffer and a disk file, and optionally in a cache. When the web crawler downloads a document from a host computer, it identifies URL's (document addresses) in the downloaded document. Each identified URL is converted into a fixed size numerical representation. The numerical representation may optionally be systematically compared to the contents of a cache containing web sites which are likely to be found during the web crawl, for example previously visited web sites. The numerical representation is then systematically compared to numerical representations in the buffer, which stores numerical representations of recently-identified URL's. If the representation is not found in the buffer, it is stored in the buffer. When the buffer is full, it is ordered and then merged with numerical representations stored, in order, in the disk file. In addition, the document corresponding to each representation not found in the disk file during the merge is scheduled for downloading. The disk file may be a sparse file, indexed to correspond to the numerical representations of the URL's, so that only a relatively small fraction of the disk file must be searched and re-written in order to merge each numerical representation in the buffer.